Webis at TREC 2014:
Web, Session, and Contextual Suggestion Tracks

Matthias Hagen Steve Göring Maximilian Michel Georg Müller Benno Stein
Bauhaus-Universität Weimar
99421 Weimar, Germany
<first name>.<last name>@uni-weimar.de

ABSTRACT

In this paper we give a brief overview of the Webis group's participation in the TREC 2014 Web, Session and Contextual Suggestion tracks. All our runs for the Web and the Session track are on the full ClueWeb12 and use the online Indri retrieval system hosted at CMU. Our runs for the Contextual Suggestion track are based on the open web.

As for the Web track, our runs are aimed at one research question: whether using axioms for re-ranking a baseline result list improve the retrieval performance. Therefore, we implement the axioms available in the axiomatic IR literature and combine them with new axioms aimed at term proximity. Trained on the TREC 2013 Web track data, three promising combinations of axioms are identified in a large-scale experiment and used for our three runs.

As for the session track, we tackle three research questions in three different runs. First, similar to the Web track, we examine whether an axiom combination helps to improve session retrieval. Second, we examine the effect of presenting relevant documents from previous years when they seem to be related to the current queries of the 2014 data. Our third question is whether the user interactions can be used to train an activation model to predict relevant documents for new queries.

As for the contextual suggestion track, our research question is whether explanations based on the user profile and explaining why specific entities are suggested by the system are perceived positively or negatively by the user. Our focus is not explicitly on finding the most relevant suggestions but rather on examining the effect of different descriptions. Thus, the system for finding the suggestions is simply based on techniques shown promising in the previous years. Our first run uses "standard" descriptions while in the second run the standard description is enriched by a profile specific sentence explaining the reasoning underlying the suggestion.

1. RETRIEVAL SYSTEM

Our runs for the Web and the Session track are on the full ClueWeb12 corpus (category A) and use the provided baseline result lists in a re-ranking approach. In case that some further documents were needed, we us the language modeling based Indri search engine provided by the Carnegie Mellon University.[1]

Our runs for the Contextual suggestion track are on the open web and use the Google Places API to identify relevant suggestions and the Yandex Rich Content API to generate the descriptions for suggested entities.

[1]http://lemurproject.org/clueweb12/services.php

2. WEB TRACK

The research question we examine in the Web track is whether axiomatic re-ranking of the baseline retrieval system's result list can improve retrieval performance.

We will first describe the basic idea underlying our axiomatic approach, then give some details of the training process that identifies promising axiom combinations, and then briefly describe the three combinations used in our three runs.

2.1 Axiomatic Re-Ranking Approach

The main idea underlying our Web track runs is to employ combinations of axioms from the IR literature to re-rank the baseline's results. We first give a high-level view of the general system and then introduce the axioms we use to generate individual re-ranked lists of the baseline's results. We then describe the rank aggregation method used to combine different re-ranked result lists.

2.1.1 *General Setup*

In the recent IR literature, several axioms are proposed that retrieval models should follow to generate good rankings; Hui Fang's web page gives a good overview of the existing literature.[2]

Our main idea is to use combinations of such axioms as the basis of retrieval models that "by definition" follow the axioms as best as possible. The respective process consists of three steps similar to a Learning-to-Rank framework (cf. Figure 1). First, an initial ranking is obtained from a baseline retrieval system (Indri in our case) from which only the top n retrieved documents are considered as the candidate set for further processing ($n = 50$ in our case). In a second step, each axiom is used to create an individual re-ranking of the candidates. Finally, in the third step the rankings of

[2]http://www.eecis.udel.edu/~hfang/AX.html

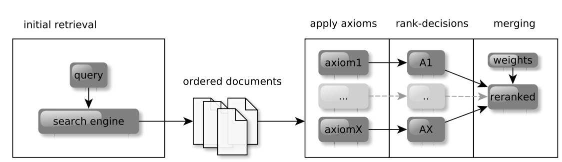

Figure 1: Our general axiomatic approach with its three steps: initial ranking, applying axioms, merge and weight results

Report Documentation Page	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE **NOV 2014**	2. REPORT TYPE	3. DATES COVERED **00-00-2014 to 00-00-2014**

4. TITLE AND SUBTITLE **Webis at TREC 2014: Web, Session, and Contextual Suggestion Tracks**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Bauhaus-Universit??t Weimar,99421 Weimar, Germany,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

presented in the proceedings of the Twenty-Third Text REtrieval Conference (TREC 2014) held in Gaithersburg, Maryland, November 19-21, 2014. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA).

14. ABSTRACT

In this paper we give a brief overview of the Webis group???s participation in the TREC 2014 Web, Session and Contextual Suggestion tracks. All our runs for the Web and the Session track are on the full ClueWeb12 and use the online Indri retrieval system hosted at CMU. Our runs for the Contextual Suggestion track are based on the open web. As for the Web track, our runs are aimed at one research question: whether using axioms for re-ranking a baseline result list improve the retrieval performance. Therefore, we implement the axioms available in the axiomatic IR literature and combine them with new axioms aimed at term proximity. Trained on the TREC 2013 Web track data three promising combinations of axioms are identified in a large-scale experiment and used for our three runs. As for the session track, we tackle three research questions in three different runs. First, similar to the Web track, we examine whether an axiom combination helps to improve session retrieval. Second, we examine the effect of presenting relevant documents from previous years when they seem to be related to the current queries of the 2014 data. Our third question is whether the user interactions can be used to train an activation model to predict relevant documents for new queries. As for the contextual suggestion track, our research question is whether explanations based on the user profile and explaining why specific entities are suggested by the system are perceived positively or negatively by the user. Our focus is not explicitly on finding the most relevant suggestions but rather on examining the effect of different descriptions. Thus, the system for finding the suggestions is simply based on techniques shown promising in the previous years. Our first run uses ???standard??? descriptions while in the second run the standard description is enriched by a profile specific sentence explaining the reasoning underlying the suggestion.

<table>
<tr><td colspan="6">15. SUBJECT TERMS</td></tr>
<tr><td colspan="3">16. SECURITY CLASSIFICATION OF:</td><td rowspan="2">17. LIMITATION OF ABSTRACT
Same as Report (SAR)</td><td rowspan="2">18. NUMBER OF PAGES
4</td><td rowspan="2">19a. NAME OF RESPONSIBLE PERSON</td></tr>
<tr><td>a. REPORT
unclassified</td><td>b. ABSTRACT
unclassified</td><td>c. THIS PAGE
unclassified</td></tr>
</table>

Table 1: **Analyzed axioms grouped by their purpose.**

Purpose	Acronym	Source	Used
Term frequency	TFC1	[8]	Yes
	TFC2	[8]	Yes
	TFC3	[9]	Yes
	TDC	[8]	Yes
Document length	LNC1	[8]	Yes
	LNC2	[8]	Yes
	TF-LNC	[8]	Yes
Lower bound	LB1	[19]	Yes
	LB2	[19]	No
Semantic similarity	STMC1	[10]	Yes
	STMC2	[10]	No
	STMC3	[10]	No
Term proximity	PROX1	new	Yes
	PROX2	new	Yes
	PROX3	new	Yes
	PROX4	new	Yes
	TSSC1/2	[20]	No
Other	R/AND	[22, 21]	No
	CPRF	[5]	No
	CTM	[17]	No
	CMR	[12]	No
	CEM	[2]	No
	ORG		Yes

the individual axioms are aggregated into a final ranking of the candidate documents.

2.1.2 Axioms for IR Models

To form a basic axiom set used in our approach, we analyzed the published axioms and selected the ones that can be used to produce re-rankings of a candidate set. We decided to only use axioms working on pairs and triples of documents. Furthermore, we decided that for the used axioms it has to be possible to formulate them as a triple

$$axiom = (precondition, condition, conclusion),$$

where *precondition* is any evaluable condition, *condition* is a more specific filter argument, and *conclusion* is a rank information (e.g., $d_i > d_j$ meaning document d_i should be ranked above document d_j). To apply an axiom, we need to iterate over all pairs or triples of candidate documents and check *precondition* and *condition* to infer the rank information.

For example, after applying axiom a we might get the ranking information $d_1 > d_2$, $d_3 > d_2$, and $d_3 > d_1$. Such results are stored in a matrix A_a for each axiom a with

$$A_a[i,j] = \begin{cases} 1 & \text{if } d_i > d_j \\ 0 & \text{otherwise} \end{cases}.$$

2.1.3 Used Axioms And Applied Modifications

The axioms that we analyzed for applicability in our setting are shown in Table 1 grouped according to their main purpose. We did not use all published axioms since not all axioms could be transformed to our needed triple form of (*precondition*, *condition*, *conclusion*), some axioms are too general or are partially covered by other axioms. For later evaluation, we also introduce an "axiom" ORG that simply yields an unmodified ranking (i.e., the original ranking).

The *precondition* of some axioms explicitly requests that a pair/triple of documents is of the exact same length (e.g., TFC3 [9]). Since in a practical setting there are hardly any documents of the exact same length, we decided to simplify such length conditions to be different by at most 10%.

Some axioms request equal term-frequency values $tf(w)$ or term-discrimination values $td(w)$ (often *idf* is used). We decided to truncate such values to the first two decimals, since otherwise there are hardly any terms with equal $tf(x)$ or $td(x)$ values.

Some axioms employ a similarity measure $sim(w_1, w_2)$ for two terms. We use WordNet[3] in these cases.

2.1.4 Rank-aggregation

Based on the ranking matrices A_a of the individual used axioms, we create a summarized and weighted matrix A with

$$A = \sum_{a \in axioms} A_a \cdot weight(a),$$

where the *weight* function ($weight : axioms \mapsto \mathbb{R}^+$) is used to increase or decrease the importance of an axiom in different settings.

Note that the matrix A might contain conflicts: if for example $A[i,j] = A[j,i]$ there is no clear answer to the question of whether document d_i should be ranked above d_j or vice versa. This describes a typical rank-aggregation problem that can be translated to a social choice problem for which several rank aggregation schemes are proposed [6, 4]. We choose the Kemeny score for rank-aggregation since it can be used in meta-search engines for rank aggregation [6] and since we can use the particular rankings of the individual axioms. Kemeny rank aggregation aggregates m different rankings into one global ranking while minimizing the sum of a distance function to the original m rankings. The distance typically denotes the number of pairs of candidates that are ranked in a different ordering [18].

Identifying the Kemeny ranking in election setups is a well known NP-complete problem [16]. Different Kemeny rank aggregation approaches are proposed in the literature for instance based on FPT-ideas [3] or the KwikSort approximation scheme [1]. We employ KwikSort that originally solves the minimum feedback arc set problem in weighted tournaments. Our setting of Kemeny rank aggregation can be transformed to such a problem by viewing it as a directed weighted graph with the vertex set $V = \{d_1, ..., d_n\}$ and edges described by the above matrix A.

2.2 Identifying Promising Combinations

Based on the described axiomatic re-ranking schemes, we develop runs for the Web Track 2014. To identify promising combinations of the individual axioms, we use the Web Track 2013 queries and relevance judgments.

For each Web Track 2013 query and for each combination of axioms (overall 2^{14} for the 14 axioms used), we build the resulting rankings. For every ranking of each combination, we calculate the difference of the $\alpha - nDCG@20$ value to the baseline. Based on these measures, we generate the two combinations "max" and "syn."

To identify the max-combination, we calculate the mean score over all queries for each axiom combination. To identify the syn-combination, we count for how many queries

[3] http://wordnet.princeton.edu/

the score difference to the baseline is negative. Using the described values calculated for all axiom combinations, we identify the candidate combinations as follows. First, we select the top performing 10% of the axiom combinations (we use 1,600 out of the 16,384). If an axiom a is contained in more than half of these combinations (i.e., more than 800) it will be used for the respective max- or syn-combination. Note that the axioms contained in a combination are weighted uniformly. Further tuning the weights might be an interesting task for future work.

2.3 Runs

Our three runs are using different axiom combinations for generating the re-ranked result lists from the top-50 documents of the baseline ranking.

- *webisWt14axAll; All axioms:* This run uses all 14 implemented axioms.
- *webisWt14axMax; Max-combination:* This run uses the axioms TFC3, LNC1, LNC2, LB1, PROX2, PROX3, and PROX4.
- *webisWt14axSyn- Synthetic syn-combination:* This run uses the axioms TFC1, TDC, LNC2, LB1, PROX2 and PROX3.

2.4 Brief Discussion

The results of our different runs indicate that the syn- and max-combinations can significantly improve the baseline ranking. Both clearly outperformed the all-combination. Thus, carefully choosing axiom combinations tailored to different retrieval models might be a very promising axiomatic re-ranking idea that we aim to further explore.

3. SESSION TRACK

The three research questions we examine with our runs of the Session track are as follows. In a first run, we use the axiomatic approach explained for the Web track to get some more judgments for evaluating the axiom combinations. In a second run, we examine whether displaying results judged as relevant for similar information needs in the last year, help to improve retrieval performance of new but related queries. In a third run, we examine whether the user interactions can be used to train an activation model to predict relevant documents for new queries.

In the following, we explain the three runs individually. For each run, three ranked lists have to be submitted. The first RL1 does not use any session information, so we use the baseline results provided by the track organizers without any changes. For the second list RL2, local session knowledge can be used, for example the displayed results and the clicks of the session. For the third list RL3, any knowledge also from other sessions might be exploited.

3.1 Run webisSt14ax

The idea of our axiomatic run for the Session track is to use the syn-combination for RL2 and the max-combination for RL3. To exploit session knowledge, we train some further weighting compared to the Web track runs where an axiom was either used or not but the used axioms had uniform weights. The axiom weights for the session track are trained on the previous clicks of the user submitting the current query. Thus, both our RL2 and RL3 of the webisSt14ax run only use "local" session information of the same user. As before, the first 50 documents of the baseline are re-ranked and the remaining baseline results are simply appended without further modification.

3.1.1 Axiom Weighting

To train the individual axiom weights in the max- and the syn-combination, we use the session information in form of previous queries, retrieved results with shown snippets, and click information. Based on the previous queries and clicked results, we interpret each click interaction for previous queries as a manual re-ranking: simply swapping the documents based on the click logs. We assume that the user just saw the snippet and thus only use the snippets for click interpretation (i.e., viewing the snippets as a compressed version of the document). The resulting user re-ranked result is used to find an axiom weight combination that yields the most similar ranking. The "best" weight combination for an individual session is then also used in the max- and syn-combination for the current query of the session.

3.1.2 Detailed RLs

As for RL1, we simply use the original baseline results. As for RL2, we use the syn-combination from the Web track with trained weights on the individual sessions. As for RL3, we use the max-combination in an analogous manner.

3.2 Run webisSt14db

The idea of our second Session track run is to examine the influence of documents judged as relevant to similar sessions / information needs. To this end, we exploit the relevance judgments from the TREC 2013 Session track and insert documents at the top of the current ranking with the highest judgments on similar sessions from 2013 if there are any that are at least relevant. Session similarity was measured using our session detection scheme [13]: only sessions from other users with a cosine similarity (*tf*-weights) of the query strings of more than 0.35 are viewed as similar. Note that this approach is somewhat similar to our last year's run webisS1 [15] but with qrel files instead of click logs (i.e., assuming a perfect knowledge of what others perceive as relevant).

3.2.1 Detailed RLs

As for RL2, we only remove documents contained in the result list of previous queries from the baseline ranking. As for RL3, we insert the documents judged as relevant or better from sessions of the previous year and then append the baseline results without the previously seen documents.

3.3 Run webisSt14act

The idea of our third run is to examine the impact of a user model inspired by the spreading activation framework [11]. Thus, for RL3 we assume the knowledge of the task description—somewhat cheating but also modeling an almost perfect task model. Based on keyphrases extracted from the task description, the activation of a document's snippet is measured.

3.3.1 Detailed RLs

RL2 is the same as for our webisSt14db run: we only remove previously seen documents from the baseline ranking.

As for RL3, we re-rank the RL2 list according to the activation based on PMI between phrases of the task description and the result snippets. Documents with higher activation snippets are moved to the top with the reasoning that the user would probably rather click on these. In case that the RL3 results are promising, a further RL2 could be trained on only the keyphrases extracted from the users previous queries, for instance via query segmentation [14], and on keyphrases extracted from the (clicked) snippets.

3.4 Brief Discussion

The results indicate that our idea of training axiom combinations based on the user clicks did not improve the ranking very much. One reason might be that only few clicks are contained in the released sessions such that is almost impossible to "train" at all. Experiments with longer sessions containing more clicks might be an interesting direction for the axiomatic re-ranking framework. As for the activation-based run, it seems to also have not helped much to improve the result ranking.

4. CONTEXTUAL SUGGESTION TRACK

The research question we examine in the Contextual Suggestion track is whether a description enriched by an explanation of why an entity was suggested is perceived as positive or negative. Thus, for the actual retrieval of suggestions we build upon state-of-the-art tools. Our first run uses descriptions without explanations while an explanation is added in the second run. The suggestions are the same for both runs, only the descriptions differ.

4.1 Run webis_1

To identify the suggestions of our first run, we crawled suggestions from the Google Places API near the context coordinates (radius 2.5 km) and rank them according to the Google Places ranking. The description for an item is generated via the Yandex Rich Content API with the respective Google Places website as input. The description is restricted to a maximum of 330 characters to be able to extend it with further information in our second run.

4.2 Run webis_2

The suggestions are the ones from the first run but the descriptions are slightly different. The 330 characters obtained from the Yandex Rich Content API are preceded by a sentence containing the average user rating from the Google Places API and whether the user for whom the suggestion is derived favors similar examples in their profile.

4.3 Brief Discussion

The results indicate that the additional information given in our second run's descriptions did not result in better judgments. Probably placing the further explanations of average ratings and similar examples in front of the actual item's description is not the best idea. Promising directions for future work might be to examine the placement of the explanation at the end of the actual description text or even somewhere in the middle. In this case of choosing to insert it into the middle, paraphrasing techniques from computational linguistics might be helpful to polish the combined description and ensure a good readability for human users.

5. REFERENCES

[1] N. Ailon, M. Charikar, and A. Newman. Aggregating inconsistent information: Ranking and clustering. *Journal of the ACM*, 55(5):23:1–23:27, 2008.

[2] E. Amigó, J. Gonzalo, and F. Verdejo. A general evaluation measure for document organization tasks. In *Proceedings of SIGIR 2013*, pages 643–652.

[3] N. Betzler, M. R. Fellows, J. Guo, R. Niedermeier, and F. A. Rosamond. Fixed-parameter algorithms for Kemeny scores. In *Theoretical Computer Science*, 410(45):4554–4570, 2009.

[4] Y. Chevaleyre, U. Endriss, J. Lang, and N. Maudet. A short introduction to computational social choice. *Proceedings of SOFSEM 2007*. pages 51-69.

[5] S. Clinchant and É. Gaussier. A document frequency constraint for pseudo-relevance feedback models. In *Proceedings of CORIA 2011*, pages 73–88.

[6] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proceedings of WWW 2001*, pages 613–622.

[7] H. Fang. A re-examination of query expansion using lexical resources. In *In Proceedings of ACL 2008*, pages 139–147.

[8] H. Fang, T. Tao, and C. Zhai. A formal study of information retrieval heuristics. In *Proceedings of SIGIR 2004*, pages 49–56.

[9] H. Fang, T. Tao, and C. Zhai. Diagnostic evaluation of information retrieval models. *ACM Transactions on Information Systems*, 29(2):7:1–7:42, 2011.

[10] H. Fang and C. Zhai. Semantic term matching in axiomatic approaches to information retrieval. In *Proceedings of SIGIR 2006*, pages 115–122.

[11] W. Fu and P. Pirolli. SNIF-ACT: A cognitive model of user navigation on the world wide web. *Human-Computer Interaction*, 22(4):355–412, 2007.

[12] S. Gerani, C. Zhai, and F. Crestani. Score transformation in linear combination for multi-criteria relevance ranking. In *Proceedings of ECIR 2012*, pages 256–267.

[13] M. Hagen, J. Gomoll, A. Beyer, and B. Stein. From search session detection to search mission detection. In *Proceedings of OAIR 2013*, pages 85–92.

[14] M. Hagen, M. Potthast, A. Beyer, and B. Stein. Towards optimum query segmentation: In doubt without. In *Proceedings of CIKM 2012*, pages 1015–1024.

[15] M. Hagen, M. Völske, J. Gomoll, M. Bornemann, L. Ganschow, F. Kneist, A. H. Sabri, and B. Stein. Webis at TREC 2013 Sessions and Web track. In *Proceedings of TREC 2013*.

[16] E. Hemaspaandra, H. Spakowski, and J. Vogel. The complexity of Kemeny elections. *Theoretical Computer Science*, 349(3):382 – 391, 2005.

[17] M. Karimzadehgan and C. Zhai. Axiomatic analysis of translation language model for information retrieval. In *Proceedings of ECIR 2012*, pages 268–280.

[18] J. G. Kemeny. Mathematics without numbers. *Daedalus*, 88(4):pp. 577–591, 1959.

[19] Y. Lv and C. Zhai. Lower-bounding term frequency normalization. In *Proceedings of CIKM 2011*, pages 7–16.

[20] T. Tao and C. Zhai. An exploration of proximity measures in information retrieval. In *Proceedings of SIGIR 2007*, pages 295–302.

[21] H. Wu and H. Fang. Relation based term weighting regularization. In *Proceedings of ECIR 2012*, pages 109–120.

[22] W. Zheng and H. Fang. Query aspect based term weighting regularization in information retrieval. In *Proceedings of ECIR 2010*, pages 344–356.